REPORT DOCUMENTATION PAGE	*Form Approved* **OMB No. 0704-0188**

1. REPORT DATE *(DD-MM-YYYY)* 05-17-2012	**2. REPORT TYPE** Final Report	**3. DATES COVERED** *(From - To)* July 2010 - June 2012

4. TITLE AND SUBTITLE Integrated Warfighter Biodefense Program (IWBP)	**5a. CONTRACT NUMBER** N00014-10-C-0363
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 0603729N
6. AUTHOR(S) Abbott, Franklin T. Vaidyanathan, Ganesh	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Quantum Leap Innovations, Inc. 3 Innovation Way, Suite 100 Newark, DE 19711-5456	**8. PERFORMING ORGANIZATION REPORT NUMBER** QLI-TR-05-001
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Office of Naval Research ONR Code 342 875 North Randolph Street Arlington, VA 22203-1995	**10. SPONSOR/MONITOR'S ACRONYM(S)** ONR
	11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
Distribution Statement A: Approved for public release; distribution is unlimited. 17 May 2012.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The aim of the Integrated Warfighter Biodefense Program (IWBP) is to develop innovative technology that can be deployed to prevent U.S. armed forces from becoming battle or non-battle casualties, and especially to reduce morbidity and mortality throughout the increasingly complex battlespace of current operations. In this summary of the next phase of work on IWBP we report on the continued development of data analysis software that addresses the complex challenges of large incomplete datasets such as those of sensor data environments, pharmacologic datasets for drug discovery, and electronic medical records targeted towards identification of optimal treatment strategies and adverse medical events. This technology allows for examination of large databases to identify relevant information that can be used as the basis for subsequent modeling and analysis. This technology supports analysis of large, diverse datasets aimed at Force Health Protection within the broader mission. Together, these tools can be used in a number of different modalities to support the broad capability of transforming increasingly more complex multi-source, multimodal data environments into actionable knowledge.

15. SUBJECT TERMS
Biological Defense, Force Health Protection, Situational Awareness, Pattern Based Discovery, Pattern Based Prediction, Data Analytics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT** UU	**18. NUMBER OF PAGES** 22	**19a. NAME OF RESPONSIBLE PERSON** Dr. Ganesh Vaidyanathan
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified			**19b. TELEPONE NUMBER (*Include area code*)** 302-894-8044

Quantum Leap Innovations, Inc.
Delaware Technology Park
3 Innovation Way, Suite 100
Newark, DE 19711

QLI-TR-05-001
April 2012

Integrated Warfighter Biodefense Program (IWBP)

Final Report

ONR Contract N00014-10-C-0363

Sponsored by the Office of Naval Research, Arlington, VA 22203

Abstract

The aim of the Integrated Warfighter Biodefense Program (IWBP) is to develop innovative technology that can be deployed to prevent U.S. armed forces from becoming battle or non-battle casualties, and especially to reduce morbidity and mortality throughout the increasingly complex battlespace of current operations. In this summary of the next phase of work on IWBP we report on the continued development of data analysis software that addresses the complex challenges of large incomplete datasets such as those of sensor data environments, pharmacologic datasets for drug discovery, and electronic medical records targeted towards identification of optimal treatment strategies and adverse medical events. This technology allows for examination of large databases to identify relevant information that can be used as the basis for subsequent modeling and analysis. This technology supports analysis of large, diverse datasets aimed at Force Health Protection within the broader mission. Together, these tools can be used in a number of different modalities to support the broad capability of transforming increasingly more complex multi-source, multimodal data environments into actionable knowledge.

Contents

List of Figures

1. SUMMARY

Executive Summary

This final technical report summarizes Quantum Leap Innovations' (QLI) accomplishments with the Integrated Warfighter Biodefense Program (IWBP) through the contract close date of June 30, 2012 on ONR Contract N00014-10-C-0363.

Summary of Accomplishments

a. Data Integration, Transformation and Reduction
b. Scalable Data Analysis
c. Information Visualization

QLI has developed several data analysis capabilities aimed at analyzing large, complex data sets. These capabilities have resulted in new approaches toward data reduction and scalable inductive and deductive reasoning. Additional examination and development of the technology will fall under the following three specific tasks:

Specific Tasks:

Task L W -3. Data Integration, Transformation and Reduction.
As data acquisition continues to expand rapidly in naval and other contexts, there is increasing need to perform principled reduction of the data into subsets that are informative against the desired objective or goal. QLI's core technical approach for this task is based on using metrics derived from information theory or other statistical means to measure the information content of individual variables or combinations of

variables within large data environments. Evaluating the importance of variables through their interactions constitutes a key aspect of our technical approach. This task will develop methods for integrating, transforming and reducing data to enriched subsets for subsequent analysis.

Task L W -4. Scalable Data Analysis.
QLI will refine and develop methods for performing scalable analysis in the form of both deductive and inductive reasoning of large, complex datasets. Our technical approach for this task is based upon the principles of distributed modeling in which, populations of smaller models are built in a computationally efficient manner and subsequently combined to produce a consensus model. This approach provides both computational efficiency as well as controlled means for tuning the final model results. Example domain applications include sensor data environments, pharmacologic datasets for drug discovery, and electronic medical records targeted towards identification of optimal treatment strategies and adverse medical events in support of Force Health Protection.

Task L W -S. Information Visualization.
In order to facilitate the transition of analysis results to actionable knowledge, a key step will be how best to present the resulting information to end users in a common operational picture. This is especially true in the case of complex end-user environments such as battlefields and maritime security. This task will broadly pursue methods for presenting analysis results to end users that can facilitate expedient actions and information sharing. An important element in this task will be to develop methods for information visualization that are easily understood by end users. To that end, this task will include general methods for linking color characteristics (I.e., hue, saturation and intensity) to statistical information to provide the end user with straightforward visualization and dissemination of information. In environments where situational awareness and timely decision making are critical for mission success, such color based information presentation can be especially powerful.

The following introduction and example summarizes the implementation of all three tasks in the context of an integrated Pattern Based Discovery and Prediction technology capability that is demonstrated in the context of two health care examples;

2. BACKGROUND - PATTERN BASED ANALYTICS

Pattern Based Analytics is a powerful new approach to data analysis. It provides unique insights into the full complexity of real world data and it does this without either requiring deep mathematical skills or by requiring heroic simplifying assumptions about the important variables at work. The Quantum Leap® Pattern Based Analytics suite of products developed by Quantum Leap Innovations enables transparent, flexible discovery, visualization and analysis of informative patterns in large, complex data environments. These characteristics empower the non-statistical subject matter expert to rapidly obtain insight into their data for discovery, forecasting and decision making.

Patterns in data are prevalent across multiple domains. For example, technical financial market analysis often uses pattern recognition to identify profitable trading opportunities. In the life sciences, patterns of multi-gene associations can provide fundamental understanding of disease mechanisms as a basis for finding cures. In marketing analysis, patterns of customer behavior are fundamental to driving strategies that are customized for different customer segments. More generally, in the real world, patterns represent complex combinations of different variables or factors that drive outcomes. Patterns are a fundamental way in which we organize our experiential knowledge as a basis for decision making. The ability to discover new patterns in data can thus provide a key edge to decision makers in an ever more competitive and fast moving world.

Quantum Leap Pattern Based Analytics (PBA) has some differentiating advantages over traditional statistical Analytics approaches. It is based on a multi-dimensional extension of Shannon Information Theory developed by Claude Shannon[1], one of the founders of modern computer science. Informative patterns comprising multiple attributes can be rapidly discovered and visualized from complex, real world data. In contrast, traditional statistical methods have difficulty in identify complex, multivariate statistical associations in a transparent manner. In addition, PBA makes no assumptions about the nature of the relationships within the data; it doesn't require (usually unrealistic) assumptions of linear behavior but can handle arbitrarily non-linear relationships. In a similar vein, patterns can be discovered from data that can have arbitrary statistical distributions, in contrast to many traditional methods that assume normal or standard "bell curve" distributions. This latter advantage can be significant in many domains such as finance and health care. For example, in health care, biases in data gathering across a population can result in non-Gaussian distributions with "fat tails" where standard statistical analysis methods could lead to incorrect conclusions.

A fundamental characteristic of PBA is ease of use. Patterns can be easily understood by the non-expert end user or decision maker. The traditional data to decision making cycle within a business environment typically involves complex data analysis performed by "quants". The results of the analysis are then summarized in the form of reports that are more easily digested by the business end user. The cycle time associated with this process can lead to costly delays in the decision making process, as well as potentially lead to some information loss during the translation of the statistics to the final report. The goal of PBA is to remove the data-quant-end user cycle and empower the business end user to directly discover informative patterns in data as a basis for more timely decision making. PBA can be used in a complementary fashion with spreadsheets to provide new capability to the end user

Although pattern discovery has been employed in data analysis, it has been traditionally relegated to very specific types of analysis. For example, it has been used to discover patterns in symbolic sequences such as DNA sequences in biology. There are significant challenges to generalizing pattern based discovery to diverse, heterogeneous data environments with the complications of missing data etc that are prevalent in the real world. In addition, perhaps related to this observation, very little effort has been expended to date on utilizing the discovered patterns as a basis for more advanced analysis such as prediction and hypothesis generation. Quantum Leap PBA is aimed at addressing these gaps in the current state of art.

A useful analogy to information discovery based on Pattern Based Analytics is the art of taking a good photograph in a complex terrain. The required actions are as follows:

a. Point the camera in the right overall direction.
b. Adjust the f-stop.
c. Zoom in with the camera to more clearly see the screen.
d. Trigger the exposure.

Following this analogy a bit further, Quantum Leap PBA enables capability similar to that of a state of the art SLR camera where all the steps involved in producing a sharp image can be automatically performed under the hood to allow even the non-expert photographer to take high quality pictures. In the world of data, our goal might be to "photograph" nuggets of data that provide insight to an end user. For example, is there data in a sales database that could help a business analyst to understand why certain customers result in losses? Or data in a health care database that could help doctors understand optimal drug combinations for specific patients?

- Pointing the camera in the right overall direction translates to identifying the most informative attributes in the data where the informative object may be hiding. This step is often called dimensionality reduction in data analysis.
- Adjusting the f-stop translates to adjusting the complexity of the patterns that we are aiming to discover: Are we looking for patterns involving two attributes, three attributes etc?
- Zooming in with the camera translates to the way that we “bin” continuous data into discrete states as a basis for discovering patterns. In this context, we note that patterns are essentially discrete in nature. They can map directly to rules or queries in the database world. If continuous data is binned at too high a resolution, corresponding to an over-zoom, the full informative data object will be missed. Conversely, if continuous data is binned at too low a resolution, corresponding to an under-zoom, the data object will dissolve into an indistinguishable blur.
- Triggering the exposure translates to filtering the data with a set of discovered patterns in order to discover the informative data object that may be embedded in the data terrain. The sharpness of the final picture, or in our case, the information richness of the data object is associated with the number of patterns we use as our final filter.

The key steps of dimensionality reduction, adjusting pattern complexity, binning of continuous data, and aggregation of discovered patterns to create a composite data filter are all performed automatically by Quantum Leap PBA to enable automated pattern discovery and subsequent data analysis. Following the camera analogy, PBA also enables manual over-rides throughout the process to include human subject matter expertise to guide discovery and analysis.

Key Differentiators of PBA versus traditional statistical methods:

a. Ability to discover multi-dimensional patterns efficiently using an extension of Shannon Information Theory.

b. Ability to deal with arbitrary data relationships, both linear and non-linear. Factor analysis assumes linear relationships for characterizing data associations. Neural networks do assume arbitrary relationships, but are black box models that are not transparent to the end user.

c. Ability to deal with arbitrary statistical data distributions. This is again a result of using Shannon Information Theory. Many statistical correlation methods implicitly assume normal or Gaussian distributions.

d. Ability to deal with significant amounts of missing data using proprietary methods developed by Quantum Leap Innovations.

Awards & Honors:

Gartner Research has recognized Quantum Leap Innovations as a Cool Vendor in Life Sciences for 2011 based on in depth evaluation of Quantum Leap Pattern Based Analytics.

References:

1. C.E. Shannon, A Mathematical Theory of Communication, Bell System Technical Journal, vol. 27, pp. 379–423, 623-656, July, October, 1948.

2. Vaidyanathan, G., InfoEvolve™ - Moving from Data to Knowledge Using Information Theory and Genetic Algorithms, Annals of the NY Academy of Sciences, 1020:227-238, 2004.

3. Simmons, K. et al, Practical Outcomes of Applying Ensemble Machine Learning Classifiers to High-Throughput Screening (HTS) Data Analysis and Screening, American Chemical Society Journal of Chemical Information and Modeling November 5, 2008.

4. Simmons, K. et al, Comparative Study of Machine Learning and Chemometric Tools for Analysis of In-Vivo High-Throughput Screening Data, American Chemical Society Journal of Chemical Information and Modeling August 6, 2008

3. CONTRACT ACTIVITIES:

QLI Contract N00014-10-C-0363	$2,987,891
Award date: 07/01/2010	
ACTUAL: Expenditures Invoiced to the Government through April 30, 2012	$2,970,978
Fee withheld pending final acceptance	$ 16,913
100%+ of Contract Value has been spent as of April 30, 2012	$2,987,978

Overrun - Actual contract costs incurred exceeded total contract NTE value by $86,131

4. 1 PATTERN BASED DISCOVERY EXAMPLE

The Quantum Leap Pattern Based Discovery ("Discovery") product automatically discovers informative patterns against a user specified query using a search based paradigm. A ranked list of informative patterns that link to associated data subsets is generated from the search and displayed. This allows the user to easily perform targeted exploration, visualization and analysis of informative data subsets rather than all the data. The following example walks the user through a Healthcare Fraud problem using Discovery.

Healthcare Fraud Example:

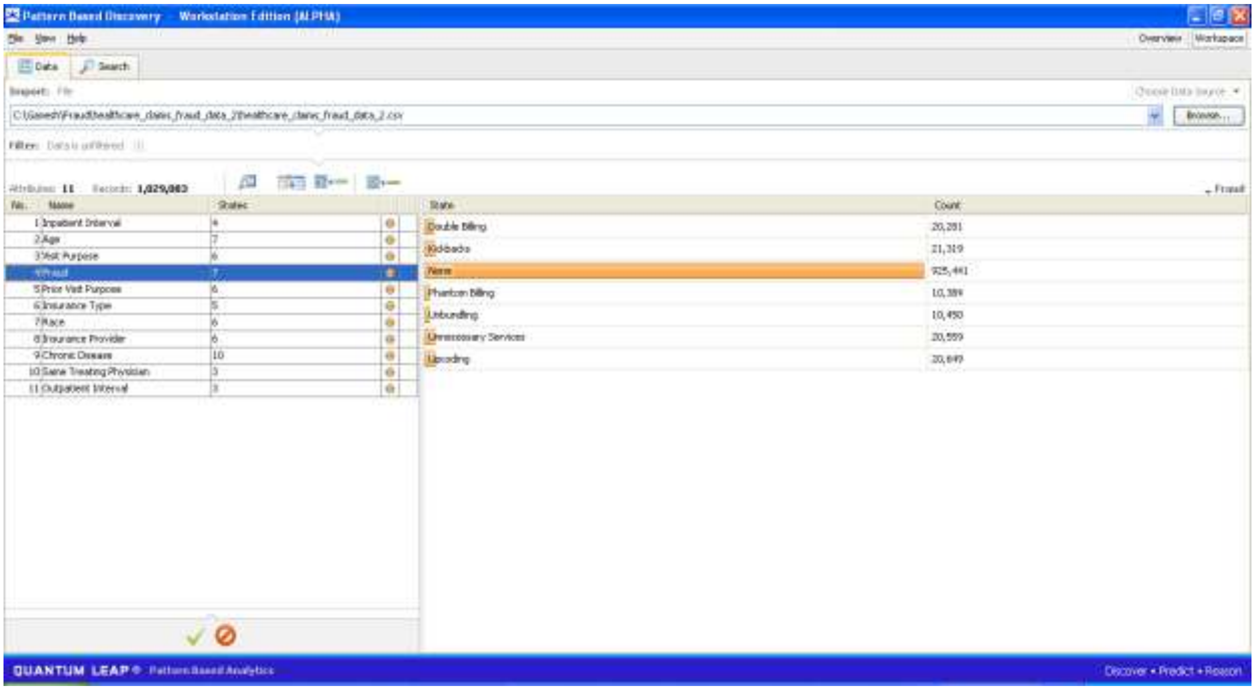
It has been estimated that healthcare fraud and abuse can constitute between 3 -15% of annual healthcare expenditures in the United States. From a cost standpoint, this translates to $100-$170 billion in annual costs! Analysis of healthcare data to discover patterns that associate with different fraud types can potentially provide a proactive means for health care providers to detect fraud early on. In this example, we use a simulated data set of ~1 million patients based on an existing fraud model ("Healthcare Fraud and Abuse", Rudman et al). Six fraud types are modeled based on statistical occurrence within the nation. An additional challenge with this data set is the prevalence of MISSING data that is characteristic of healthcare data. Appendix A summarizes the data characteristics for this example.

Key questions to be answered include:

a. What are the strongest patterns that associate with each type of Fraud?
b. Are there informative statistics/clusters within the data subsets associated with the strongest patterns that can be used as a basis for proactive monitoring of fraud?

In the following tutorial, we walk the user through the use of the Discovery product to address these types of questions.

4.1.1 Load Data ("healthcare_claims_fraud_data_2.csv") into Discovery:


Pattern Based Discovery – Workstation Edition (ALPHA)
File View Help
Overview Workspace
Data
Search
Choose Data Source
Browse...
QUANTUM LEAP Pattern Based Analytics
Discover • Predict • Reason

Import: File

C:\Ganesh\Fraud\healthcare_claims_fraud_data_2\healthcare_claims_fraud_data_2.csv

Filter: Data is unfiltered

Attributes: **11** Records: **1,029,083**

No.	Name	States
1	Inpatient Interval	4
2	Age	7
3	Visit Purpose	6
4	Fraud	7
5	Prior Visit Purpose	6
6	Insurance Type	5
7	Race	6
8	Insurance Provider	6
9	Chronic Disease	10
10	Same Treating Physician	3
11	Outpatient Interval	3

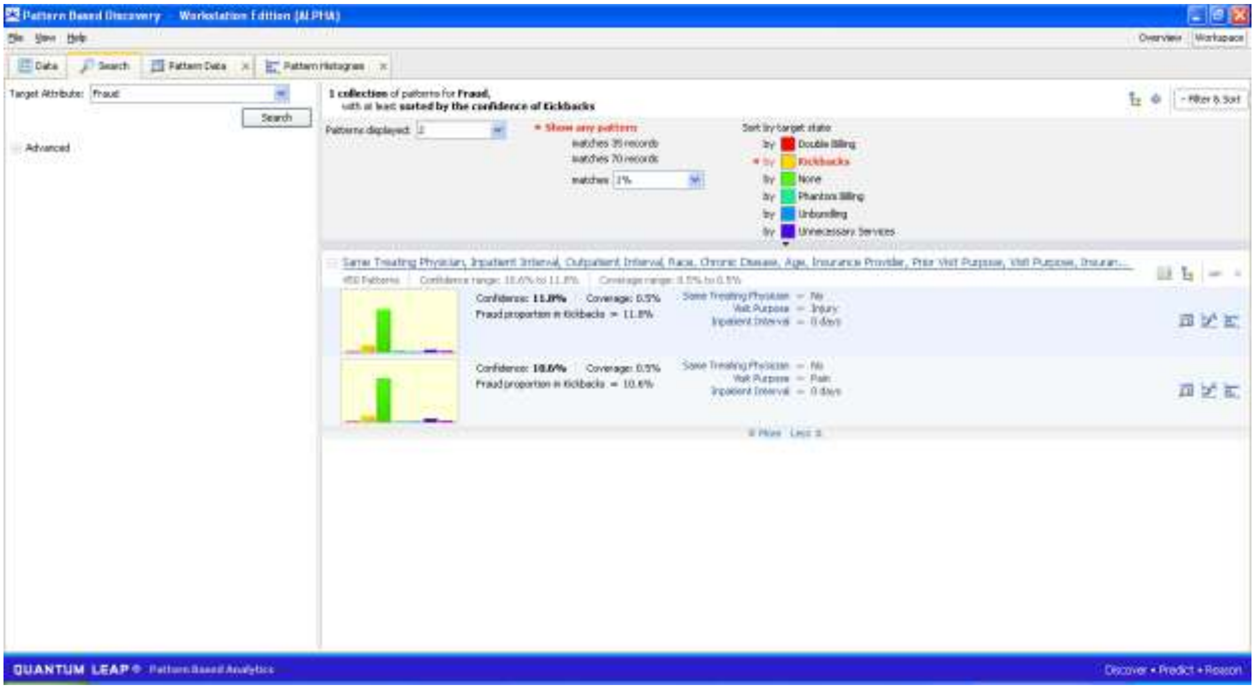
Fraud

State	Count
Double Billing	20,281
Kickbacks	21,319
None	925,441
Phantom Billing	10,389
Unbundling	10,450
Unnecessary Services	20,559
Upcoding	20,649

Figure 4.1.1 Load Data

The attributes that make up the data are shown on the left. The "Fraud" attribute is highlighted and a histogram of the corresponding distribution of fraud types is displayed on the right.

4.1.2 Search for patterns against FRAUD:


Pattern Based Discovery – Workstation Edition (ALPHA)
File View Help
Overview Workspace
Data
Search
Pattern Data
Pattern Histograms
Search
Advanced
Filter & Sort
More Less
QUANTUM LEAP Pattern Based Analytics
Discover • Predict • Reason

Target Attribute: Fraud

1 collection of patterns for **Fraud**,
with at least **sorted by the confidence of Kickbacks**

Patterns displayed: 2

● Show any pattern
matches 35 records
matches 70 records
matches 1%

Sort by target state
by 🟥 Double Billing
● by 🟨 Kickbacks
by 🟩 None
by 🟩 Phantom Billing
by 🟦 Unbundling
by 🟪 Unnecessary Services

Same Treating Physician, Inpatient Interval, Outpatient Interval, Race, Chronic Disease, Age, Insurance Provider, Prior Visit Purpose, Visit Purpose, Insuran...

450 Patterns | Confidence range: 10.6% to 11.8% | Coverage range: 0.5% to 0.5%

Confidence: **11.8%** Coverage: 0.5%
Fraud proportion in Kickbacks = 11.8%
Same Treating Physician = No
Visit Purpose = Injury
Inpatient Interval = 0 days

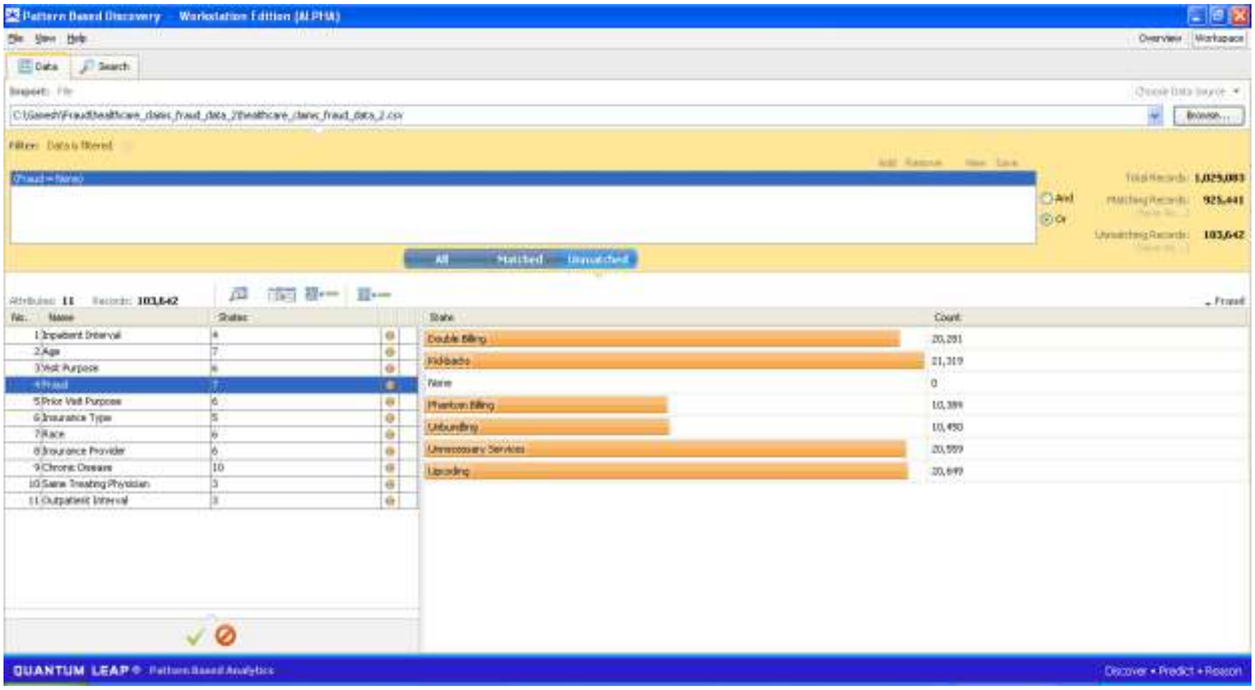
Confidence: **10.6%** Coverage: 0.5%
Fraud proportion in Kickbacks = 10.6%
Same Treating Physician = No
Visit Purpose = Pain
Inpatient Interval = 0 days

Figure 4.1.2 Search for Patterns

The search term in the top left of the screen is "Fraud". The top two patterns associated with "Kickbacks" are displayed on the right with the confidence level for Kickbacks shown as 11.8% next to the bar graph. Note that the dominant confidence level is associated with Fraud type “None” shown in green in the bar graph accompanying the pattern. This is due to the dominance of “None” within the data.

In order to “zoom in” on actual Fraud patterns, it will be useful to filter out the dominant Fraud type of “None”. We can do this by returning to the Data screen to apply a filter.

4.1.3 Go back to Data screen and filter data to exclude Fraud type “None”:


Pattern Based Discovery - Workstation Edition (ALPHA)
File View Help
Overview Workspace
Data Search
Choose Data Source
Browse...
Add Remove View Save
All Matched Unmatched
Fraud
QUANTUM LEAP Pattern Based Analytics
Discover • Predict • Reason

Import: File

C:\Saeed\Fraud\healthcare_claims_fraud_data_2\healthcare_claims_fraud_data_2.csv

Filter: Data is filtered

(Fraud = None)

○ And
● Or

Total Records: 1,029,083
Matching Records: 925,441
Unmatching Records: 103,642

Attributes: 11 Records: 103,642

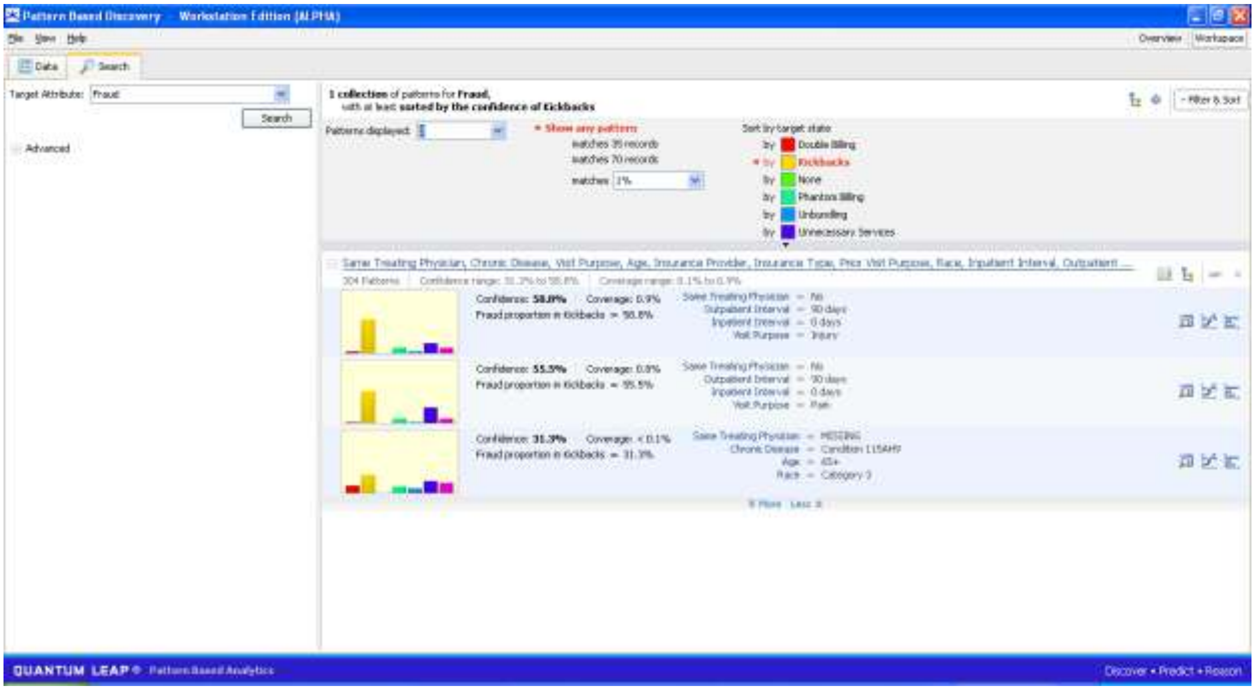
No.	Name	States
1	Inpatient Interval	4
2	Age	7
3	Visit Purpose	6
4	Fraud	7
5	Prior Visit Purpose	6
6	Insurance Type	5
7	Race	6
8	Insurance Provider	6
9	Chronic Disease	10
10	Same Treating Physician	3
11	Outpatient Interval	3

State	Count
Double Billing	20,281
Kickbacks	21,319
None	0
Phantom Billing	10,384
Unbundling	10,450
Unnecessary Services	20,559
Upcoding	20,649

Figure 4.1.3 Filter Fraud type “None”

Note the highlighted Filter (Fraud = None) within the Filter Window and the reduced number of data records (103642 records versus the original data size of 1,029,083 records) that remain after applying the filter. This filter was added by clicking on the “Add” button within the Filter window.

4.1.4 Search the filtered Fraud data to discover patterns against "Fraud":


Pattern Based Discovery – Workstation Edition (ALPHA)
File View Help
Overview Workspace
Data Search
Search
Advanced
- Filter & Sort
QUANTUM LEAP Pattern Based Analytics
Discover • Predict • Reason
More Less

Target Attribute: Fraud

1 collection of patterns for Fraud,
with at least sorted by the confidence of Kickbacks

Patterns displayed: 1

* Show any pattern
matches 35 records
matches 70 records
matches 1%

Sort by target state
by 🟥 Double Billing
* by 🟨 Kickbacks
by 🟩 None
by 🟩 Phantom Billing
by 🟦 Unbundling
by 🟪 Unnecessary Services

Same Treating Physician, Chronic Disease, Visit Purpose, Age, Insurance Provider, Insurance Type, Prior Visit Purpose, Race, Inpatient Interval, Outpatient ...

304 Patterns | Confidence range: 31.3% to 58.8% | Coverage range: 0.1% to 0.9%

Confidence: 58.8% Coverage: 0.9%
Fraud proportion in Kickbacks = 58.8%
Same Treating Physician = No
Outpatient Interval = 90 days
Inpatient Interval = 0 days
Visit Purpose = Injury

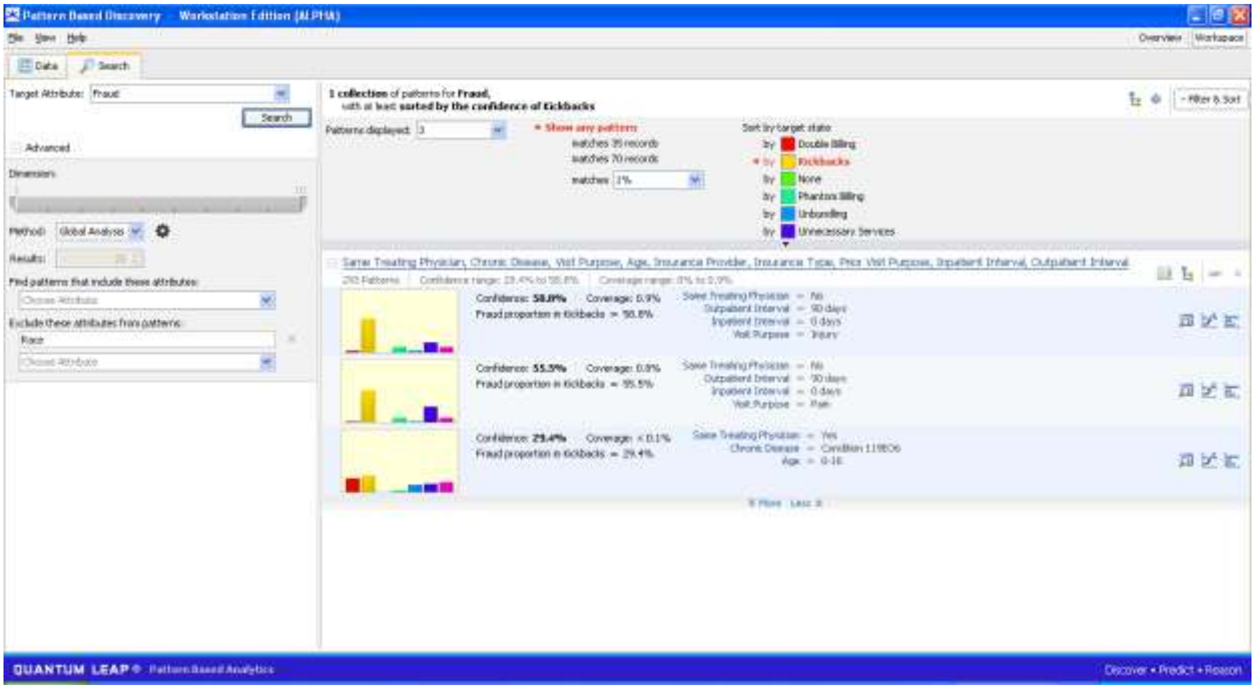
Confidence: 55.5% Coverage: 0.8%
Fraud proportion in Kickbacks = 55.5%
Same Treating Physician = No
Outpatient Interval = 30 days
Inpatient Interval = 0 days
Visit Purpose = Pain

Confidence: 31.3% Coverage: < 0.1%
Fraud proportion in Kickbacks = 31.3%
Same Treating Physician = MISSING
Chronic Disease = Condition 115AH9
Age = 65+
Race = Category 3

Figure 4.1.4 Search filtered Fraud data

We note that the confidence level for Fraud type "Kickbacks" has now increased dramatically to 58.8%. Another interesting observation is the inclusion of the attribute "Race" in the third pattern. To exclude Race as an attribute for Pattern Discovery, we can perform an Advanced Search where we can customize or refine our search.

4.1.5 Enter "Advanced" Search to exclude Race from pattern discovery:


Pattern Based Discovery – Workstation Edition (ALPHA)
File View Help
Overview Workspace
Data Search
Search
Advanced
Choose Attribute
- Filter & Sort
QUANTUM LEAP Pattern Based Analytics
Discover • Predict • Reason
More Less

Target Attribute: Fraud

Dimension:

Method: Global Analysis

Results:

Find patterns that include these attributes:

Exclude these attributes from patterns:
Race

1 collection of patterns for Fraud,
with at least sorted by the confidence of Kickbacks

Patterns displayed: 3

* Show any pattern
matches 35 records
matches 70 records
matches 1%

Sort by target state
by 🟥 Double Billing
* by 🟨 Kickbacks
by 🟩 None
by 🟩 Phantom Billing
by 🟦 Unbundling
by 🟪 Unnecessary Services

Same Treating Physician, Chronic Disease, Visit Purpose, Age, Insurance Provider, Insurance Type, Prior Visit Purpose, Inpatient Interval, Outpatient Interval

293 Patterns | Confidence range: 23.4% to 58.8% | Coverage range: 0% to 0.9%

Confidence: 58.8% Coverage: 0.9%
Fraud proportion in Kickbacks = 58.8%
Same Treating Physician = No
Outpatient Interval = 90 days
Inpatient Interval = 0 days
Visit Purpose = Injury

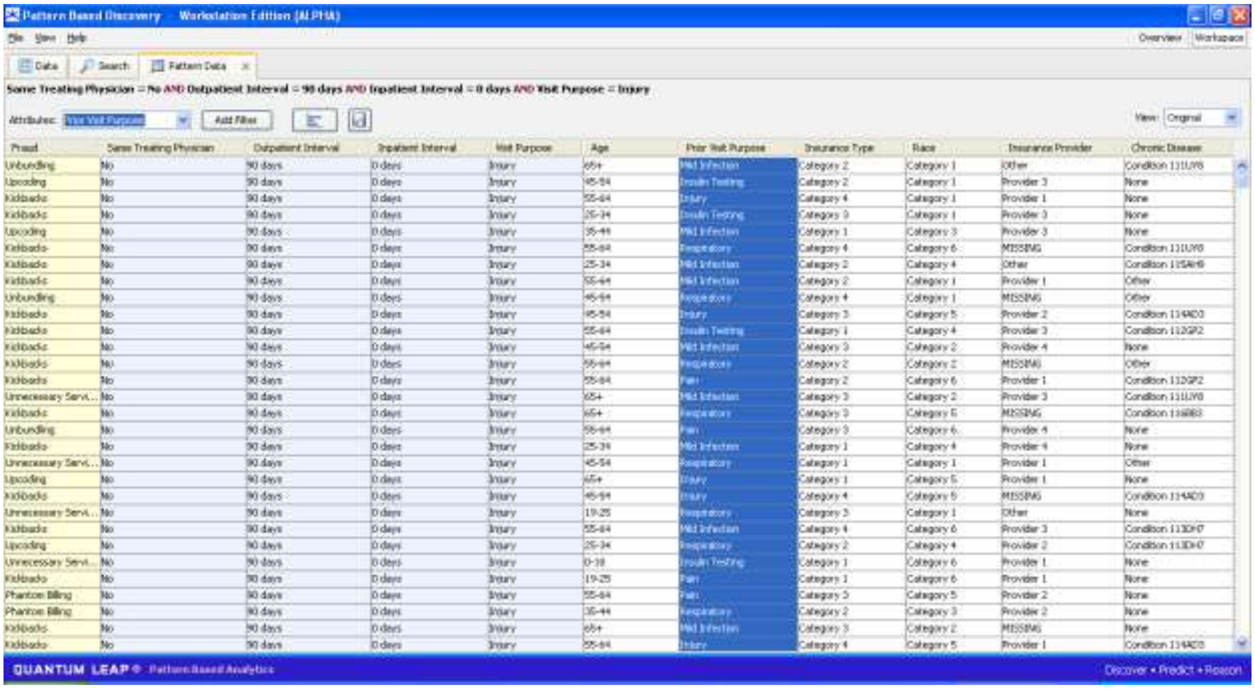
Confidence: 55.5% Coverage: 0.8%
Fraud proportion in Kickbacks = 55.5%
Same Treating Physician = No
Outpatient Interval = 30 days
Inpatient Interval = 0 days
Visit Purpose = Pain

Confidence: 29.4% Coverage: < 0.1%
Fraud proportion in Kickbacks = 29.4%
Same Treating Physician = Yes
Chronic Disease = Condition 119B06
Age = 0-18

Figure 4.1.5 Advanced Search to exclude Race

The excluded attribute "Race" is listed under the Advanced window on the lower left. Note that the resulting patterns on the right no longer include the Race attribute. We can now click on the “Show data table” icon on the right of the visual summary of each pattern to examine the data associated with the top pattern in more detail:

4.1.6 Examine data table associated with top pattern for examining the “Prior Visit Purpose” distribution within this pattern:


Pattern Based Discovery - Workstation Edition (ALPHA)
File View Help
Overview Workspace
Data
Search
Pattern Data
Add Filter
QUANTUM LEAP® Pattern Based Analytics
Discover • Predict • Reason

Same Treating Physician = No AND Outpatient Interval = 90 days AND Inpatient Interval = 0 days AND Visit Purpose = Injury

Attributes: Prior Visit Purpose

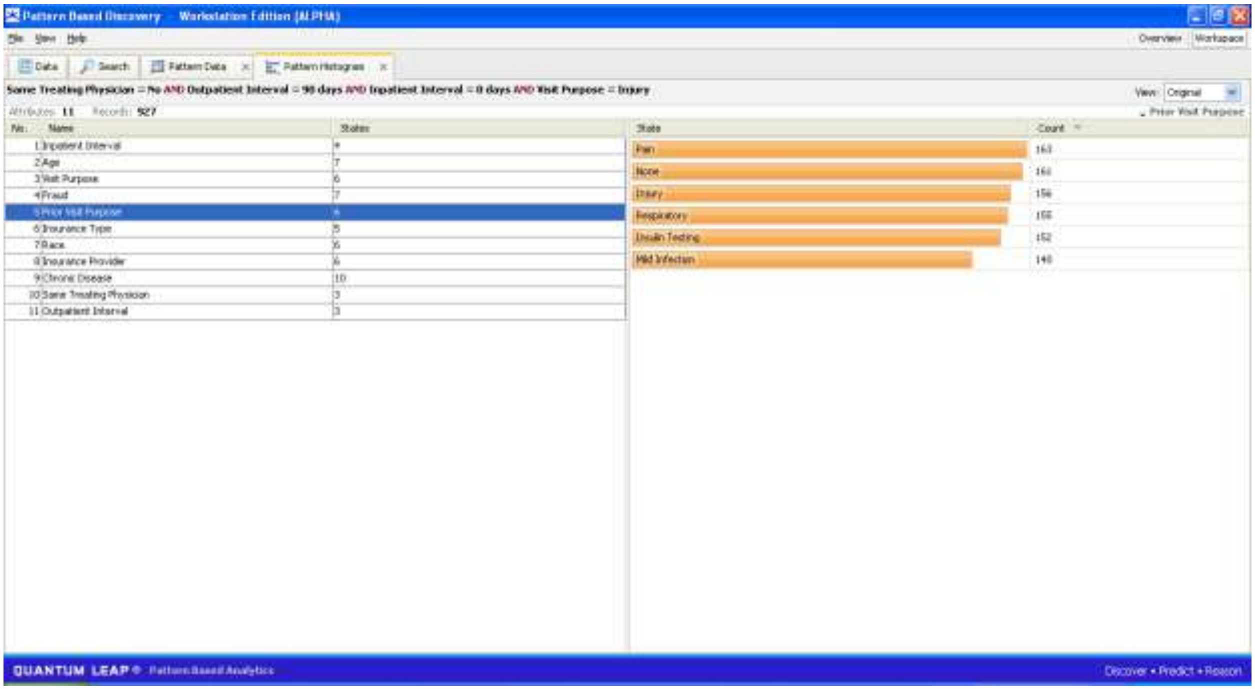
View: Original

Fraud	Same Treating Physician	Outpatient Interval	Inpatient Interval	Visit Purpose	Age	Prior Visit Purpose	Insurance Type	Race	Insurance Provider	Chronic Disease
Unbundling	No	90 days	0 days	Injury	65+	Mild Infection	Category 2	Category 1	Other	Condition 111UY8
Upcoding	No	90 days	0 days	Injury	45-54	Insulin Testing	Category 2	Category 1	Provider 3	None
Kickbacks	No	90 days	0 days	Injury	55-64	Injury	Category 4	Category 1	Provider 1	None
Kickbacks	No	90 days	0 days	Injury	25-34	Insulin Testing	Category 3	Category 1	Provider 3	None
Upcoding	No	90 days	0 days	Injury	35-44	Mild Infection	Category 1	Category 3	Provider 3	None
Kickbacks	No	90 days	0 days	Injury	55-64	Respiratory	Category 4	Category 6	MISSING	Condition 111UY8
Kickbacks	No	90 days	0 days	Injury	25-34	Mild Infection	Category 2	Category 4	Other	Condition 115AH8
Kickbacks	No	90 days	0 days	Injury	55-64	Mild Infection	Category 2	Category 1	Provider 1	Other
Unbundling	No	90 days	0 days	Injury	45-54	Respiratory	Category 4	Category 1	MISSING	Other
Kickbacks	No	90 days	0 days	Injury	45-54	Injury	Category 3	Category 5	Provider 2	Condition 114AD3
Kickbacks	No	90 days	0 days	Injury	55-64	Insulin Testing	Category 1	Category 4	Provider 3	Condition 112GP2
Kickbacks	No	90 days	0 days	Injury	45-54	Mild Infection	Category 3	Category 2	Provider 4	None
Kickbacks	No	90 days	0 days	Injury	55-64	Respiratory	Category 2	Category 2	MISSING	Other
Kickbacks	No	90 days	0 days	Injury	55-64	Pain	Category 2	Category 6	Provider 1	Condition 112GP2
Unnecessary Servi...	No	90 days	0 days	Injury	65+	Mild Infection	Category 3	Category 2	Provider 3	Condition 111UY8
Kickbacks	No	90 days	0 days	Injury	65+	Respiratory	Category 3	Category 5	MISSING	Condition 116BB3
Unbundling	No	90 days	0 days	Injury	55-64	Pain	Category 3	Category 6	Provider 4	None
Kickbacks	No	90 days	0 days	Injury	25-34	Mild Infection	Category 1	Category 4	Provider 4	None
Unnecessary Servi...	No	90 days	0 days	Injury	45-54	Respiratory	Category 1	Category 1	Provider 1	Other
Upcoding	No	90 days	0 days	Injury	65+	Injury	Category 1	Category 5	Provider 1	None
Kickbacks	No	90 days	0 days	Injury	45-54	Injury	Category 4	Category 5	MISSING	Condition 114AD3
Unnecessary Servi...	No	90 days	0 days	Injury	19-25	Respiratory	Category 3	Category 1	Other	None
Kickbacks	No	90 days	0 days	Injury	55-64	Mild Infection	Category 4	Category 6	Provider 3	Condition 113DH7
Upcoding	No	90 days	0 days	Injury	25-34	Respiratory	Category 2	Category 4	Provider 2	Condition 113DH7
Unnecessary Servi...	No	90 days	0 days	Injury	0-18	Insulin Testing	Category 1	Category 6	Provider 1	None
Kickbacks	No	90 days	0 days	Injury	19-25	Pain	Category 1	Category 6	Provider 1	None
Phantom Billing	No	90 days	0 days	Injury	55-64	Pain	Category 3	Category 5	Provider 2	None
Phantom Billing	No	90 days	0 days	Injury	35-44	Respiratory	Category 2	Category 3	Provider 2	None
Kickbacks	No	90 days	0 days	Injury	65+	Mild Infection	Category 3	Category 2	MISSING	None
Kickbacks	No	90 days	0 days	Injury	55-64	Injury	Category 4	Category 5	Provider 1	Condition 114AD3

Figure 4.1.6 Examine Data Table

The data table shows the target attribute (“Fraud”) on the far left, highlighted in yellow. The attributes that make up this pattern are shown in light blue, followed by the remaining data associated with the pattern. The selected “Prior Visit Purpose” attribute is highlighted in dark blue. A histogram of the “Prior Visit Purpose” distribution for the data described by this pattern can be displayed by clicking on the Histogram icon next to the “Add Filter” button on the top left. The data table and the pattern can be saved by clicking on the Save Data tab immediately to the right of the Histogram icon.

4.1.7 Plot Histogram of “Prior Visit Purpose” Distribution:



Pattern Based Discovery – Workstation Edition (ALPHA)

Same Treating Physician = No AND Outpatient Interval = 90 days AND Inpatient Interval = 0 days AND Visit Purpose = Injury

View: Original

Attributes: 11 Records: 927

No.	Name	States
1	Inpatient Interval	4
2	Age	7
3	Visit Purpose	6
4	Fraud	7
5	Prior Visit Purpose	6
6	Insurance Type	5
7	Race	6
8	Insurance Provider	6
9	Chronic Disease	10
10	Same Treating Physician	3
11	Outpatient Interval	3

Prior Visit Purpose

State	Count
Pain	163
None	161
Injury	156
Respiratory	155
Insulin Testing	152
Mild Infection	140

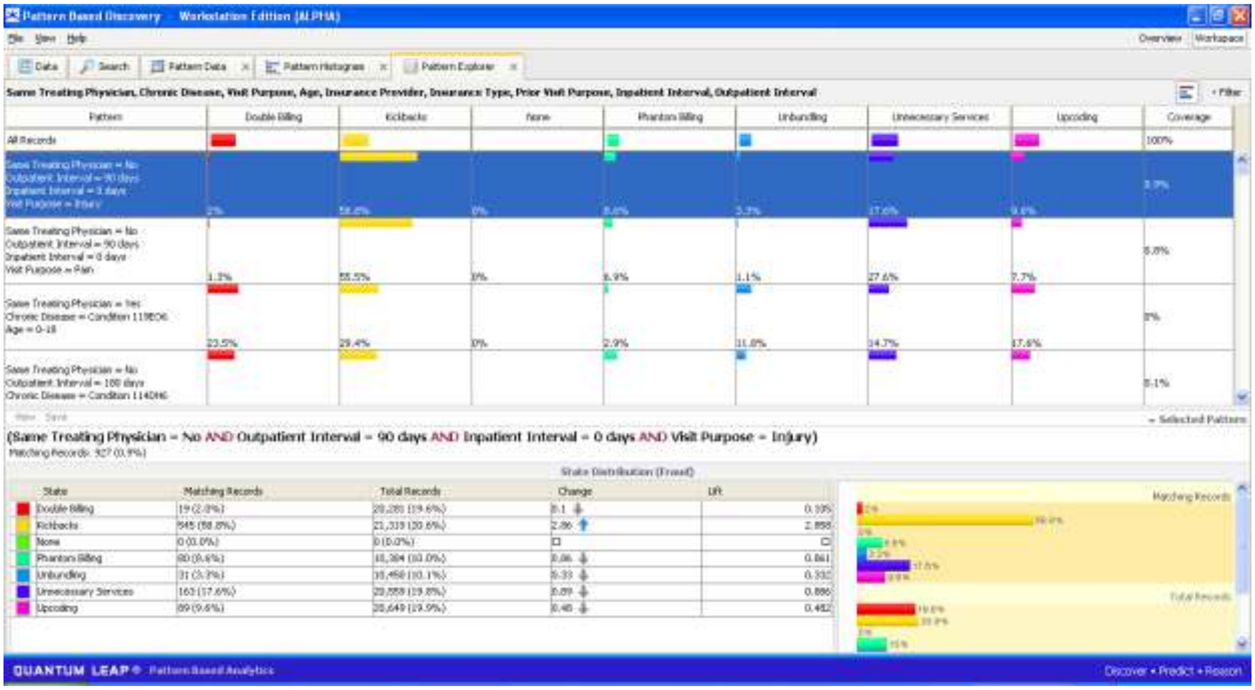
File View Help
Overview Workspace
Data Search Pattern Data Pattern Histogram
QUANTUM LEAP® Pattern Based Analytics
Discover • Predict • Reason

Figure 4.1.7 Plot Histogram

The histogram shows that the most frequently occurring “Prior Visit Purpose” category within this pattern is “Pain”.

We may further be interested in examining all 293 patterns in (E) to get a global understanding of the patterns. We can visually examine all the patterns in the collection using Pattern Explorer.

4.1.8 Visually examine all 293 patterns using Pattern Explorer:



Pattern Based Discovery – Workstation Edition (ALPHA)

Same Treating Physician, Chronic Disease, Visit Purpose, Age, Insurance Provider, Insurance Type, Prior Visit Purpose, Inpatient Interval, Outpatient Interval

Pattern	Double Billing	Kickbacks	None	Phantom Billing	Unbundling	Unnecessary Services	Upcoding	Coverage
All Records								100%
Same Treating Physician = No Outpatient Interval = 90 days Inpatient Interval = 0 days Visit Purpose = Injury	2%	58.8%	0%	8.6%	3.3%	17.6%	9.6%	0.9%
Same Treating Physician = No Outpatient Interval = 90 days Inpatient Interval = 0 days Visit Purpose = Pain	1.3%	55.5%	0%	6.9%	1.1%	27.6%	7.7%	0.8%
Same Treating Physician = Yes Chronic Disease = Condition 119E06 Age = 0-18	23.5%	29.4%	0%	2.9%	11.8%	14.7%	17.6%	0%
Same Treating Physician = No Outpatient Interval = 180 days Chronic Disease = Condition 114D46								0.1%

(Same Treating Physician = No AND Outpatient Interval = 90 days AND Inpatient Interval = 0 days AND Visit Purpose = Injury)

Matching Records: 927 (0.9%)

State Distribution (Fraud)

	State	Matching Records	Total Records	Change	Lift
🟥	Double Billing	19 (2.0%)	20,281 (19.6%)	0.1 ↓	0.105
🟨	Kickbacks	545 (58.8%)	21,319 (20.6%)	2.86 ↑	2.858
🟩	None	0 (0.0%)	0 (0.0%)	□	□
🟩	Phantom Billing	80 (8.6%)	10,384 (10.0%)	0.86 ↓	0.861
🟦	Unbundling	31 (3.3%)	10,450 (10.1%)	0.33 ↓	0.332
🟪	Unnecessary Services	163 (17.6%)	20,559 (19.8%)	0.89 ↓	0.886
🟪	Upcoding	89 (9.6%)	20,649 (19.9%)	0.48 ↓	0.482

File View Help
Overview Workspace
Data Search Pattern Data Pattern Histogram Pattern Explorer
Filter
Selected Pattern
Matching Records
Total Records
QUANTUM LEAP® Pattern Based Analytics
Discover • Predict • Reason

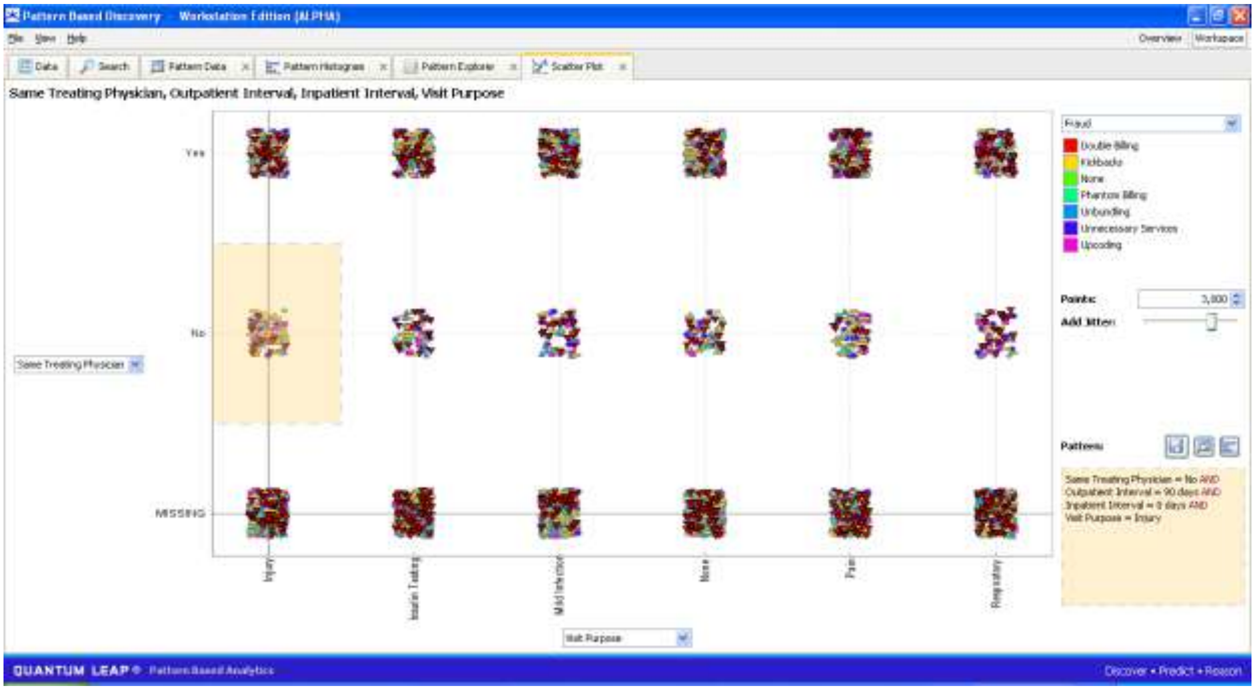
Figure 4.1.8 Explore Patterns using Pattern Explorer

Pattern Explorer aggregates all the patterns in the collection for global pattern analysis. In this example, the patterns have been sorted in descending order by the maximum confidence level for “Kickbacks”. Summary statistics on the highlighted pattern are displayed at the bottom.

The user can further examine specific relationships in data described by a pattern using the scatter plot feature where any two attributes can be plotted against each other. We demonstrate this feature using the top pattern displayed in (E).

4.1.9 Displaying x-y relationships using a scatter plot:

The top pattern in (E) involves both “Visit Purpose” and “Same Treating Physician”. When the scatter plot icon is clicked and adjusted, the user sees the scatter plot below where the user can select the attributes to be displayed from those that define the pattern and the query. In this example, we display “Same Treating Physician” versus “Visit Purpose”. The shaded yellow rectangle indicates the portion of the entire data described by the selected pattern. Note that the selected data subset shows a greater density of yellow icons representing “Kickbacks”. For discrete data, we have added “jitter” as can be seen at the right to separate overlying data. In addition, we randomly sampled 3000 data points to reduce data density.



Same Treating Physician, Outpatient Interval, Inpatient Interval, Visit Purpose

Same Treating Physician

Visit Purpose

Fraud

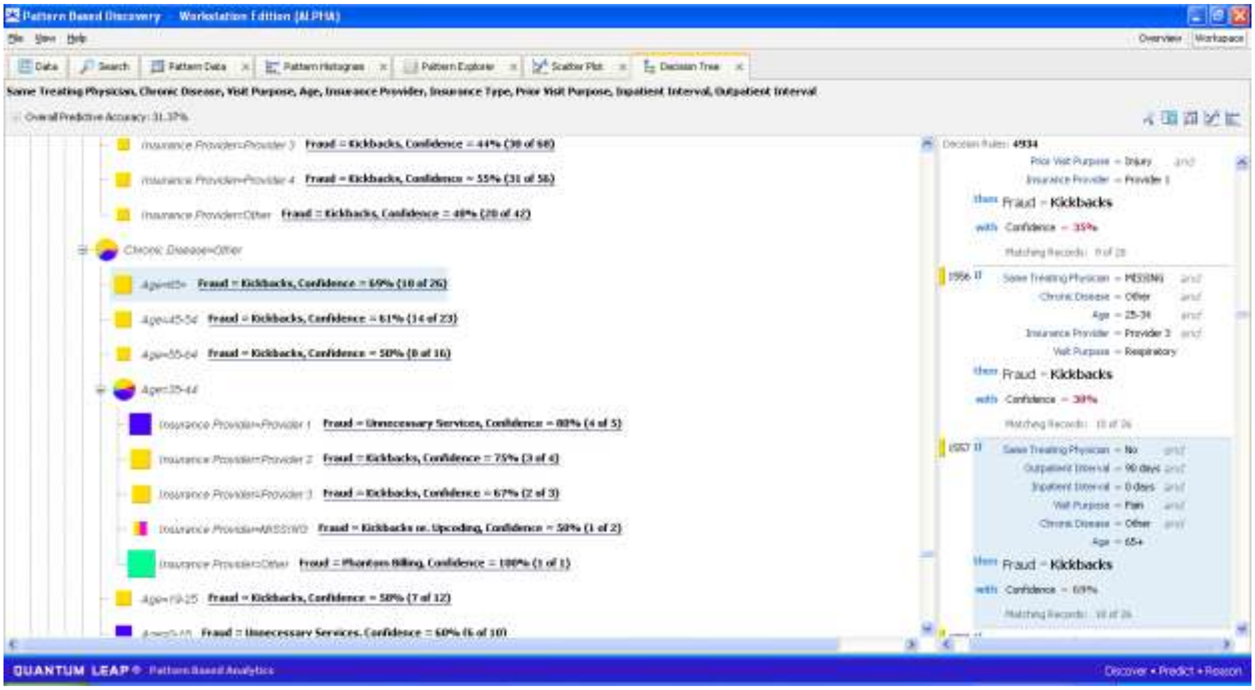
Points: 3,000

Same Treating Physician = No AND
Outpatient Interval = 90 days AND
Inpatient Interval = 0 days AND
Visit Purpose = Injury

Figure 4.1.9 Display x-y relationships using scatter plot

Finally, the user may be interested in examining more detailed relationships between all the patterns associated with the attributes that form the collection of patterns. We can “zoom in” on patterns involving this collection in more detail using the Decision Tree tab associated with the collection of patterns.

4.1.10 Displaying Inter-Pattern relationships using a Decision Tree:


Pattern Based Discovery – Workstation Edition (ALPHA)
File View Help
Overview Workspace
Data
Search
Pattern Data
Pattern Histogram
Pattern Explorer
Scatter Plot
Decision Tree
QUANTUM LEAP® Pattern Based Analytics
Discover • Predict • Reason

Same Treating Physician, Chronic Disease, Visit Purpose, Age, Insurance Provider, Insurance Type, Prior Visit Purpose, Inpatient Interval, Outpatient Interval

Overall Predictive Accuracy: 31.37%

Insurance Provider=Provider 3 Fraud = Kickbacks, Confidence = 44% (30 of 68)
Insurance Provider=Provider 4 Fraud = Kickbacks, Confidence = 55% (31 of 56)
Insurance Provider=Other Fraud = Kickbacks, Confidence = 48% (20 of 42)
Chronic Disease=Other
Age=65+ Fraud = Kickbacks, Confidence = 69% (18 of 26)
Age=45-54 Fraud = Kickbacks, Confidence = 61% (14 of 23)
Age=55-64 Fraud = Kickbacks, Confidence = 50% (8 of 16)
Age=35-44
Insurance Provider=Provider 1 Fraud = Unnecessary Services, Confidence = 80% (4 of 5)
Insurance Provider=Provider 2 Fraud = Kickbacks, Confidence = 75% (3 of 4)
Insurance Provider=Provider 3 Fraud = Kickbacks, Confidence = 67% (2 of 3)
Insurance Provider=MISSING Fraud = Kickbacks or. Upcoding, Confidence = 50% (1 of 2)
Insurance Provider=Other Fraud = Phantom Billing, Confidence = 100% (1 of 1)
Age=19-25 Fraud = Kickbacks, Confidence = 58% (7 of 12)
Age=25-34 Fraud = Unnecessary Services, Confidence = 60% (6 of 10)

Decision Rules: 4934

Prior Visit Purpose = Injury and
Insurance Provider = Provider 1
then Fraud = Kickbacks
with Confidence = 35%
Matching Records: 8 of 23

1556 If Same Treating Physician = MISSING and
Chronic Disease = Other and
Age = 25-34 and
Insurance Provider = Provider 3 and
Visit Purpose = Respiratory
then Fraud = Kickbacks
with Confidence = 38%
Matching Records: 10 of 26

1557 If Same Treating Physician = No and
Outpatient Interval = 90 days and
Inpatient Interval = 0 days and
Visit Purpose = Pain and
Chronic Disease = Other and
Age = 65+
then Fraud = Kickbacks
with Confidence = 69%
Matching Records: 18 of 26

Figure 4.1.10 Display inter-pattern relationships using Decision Tree

When the user clicks on the Decision Tree tab associated with the collection, a detailed tree is generated. Note that the highlighted pattern is even stronger than the best pattern from our search list in (g)! This is because the resolution of this local decision tree was set to a very high level using the "Pruning" icon on the top right of the screen.

This example highlights how the user can explore a collection of attributes in more detail to reveal further insights using a local decision tree whose resolution can be controlled by the user.

NOTE: The visualizations throughout the example can be saved to clipboard by right clicking on the visual. Items of interest can then be copied for example to Word to generate a report.

Summary of Data Characteristics

The states associated with each attribute can be examined using the data table in Screenshot 1.

Attributes

Inpatient Interval
Age
Visit Purpose
Prior Visit Purpose
Insurance Type
Race
Insurance Provider
Chronic Disease
Same Treating Physician
Outpatient Interval
Fraud (Double Billing, Kickbacks, Phantom Billing, Unbundling, Unnecessary Services, Upcoding)

4.2 EXTENSIONS TO PATTERN BASED DISCOVERY

The Quantum Leap Pattern Based Discovery ("Discovery") product automatically discovers informative patterns against a user specified query using a search based paradigm. A ranked list of informative patterns that link to associated data subsets is generated from the search and displayed. This allows the user to easily perform targeted exploration, visualization and analysis of informative data subsets rather than all the data. Extensions to the core capabilities of Discovery include:

4.2.1 Addition of the Jasper report generation capability:

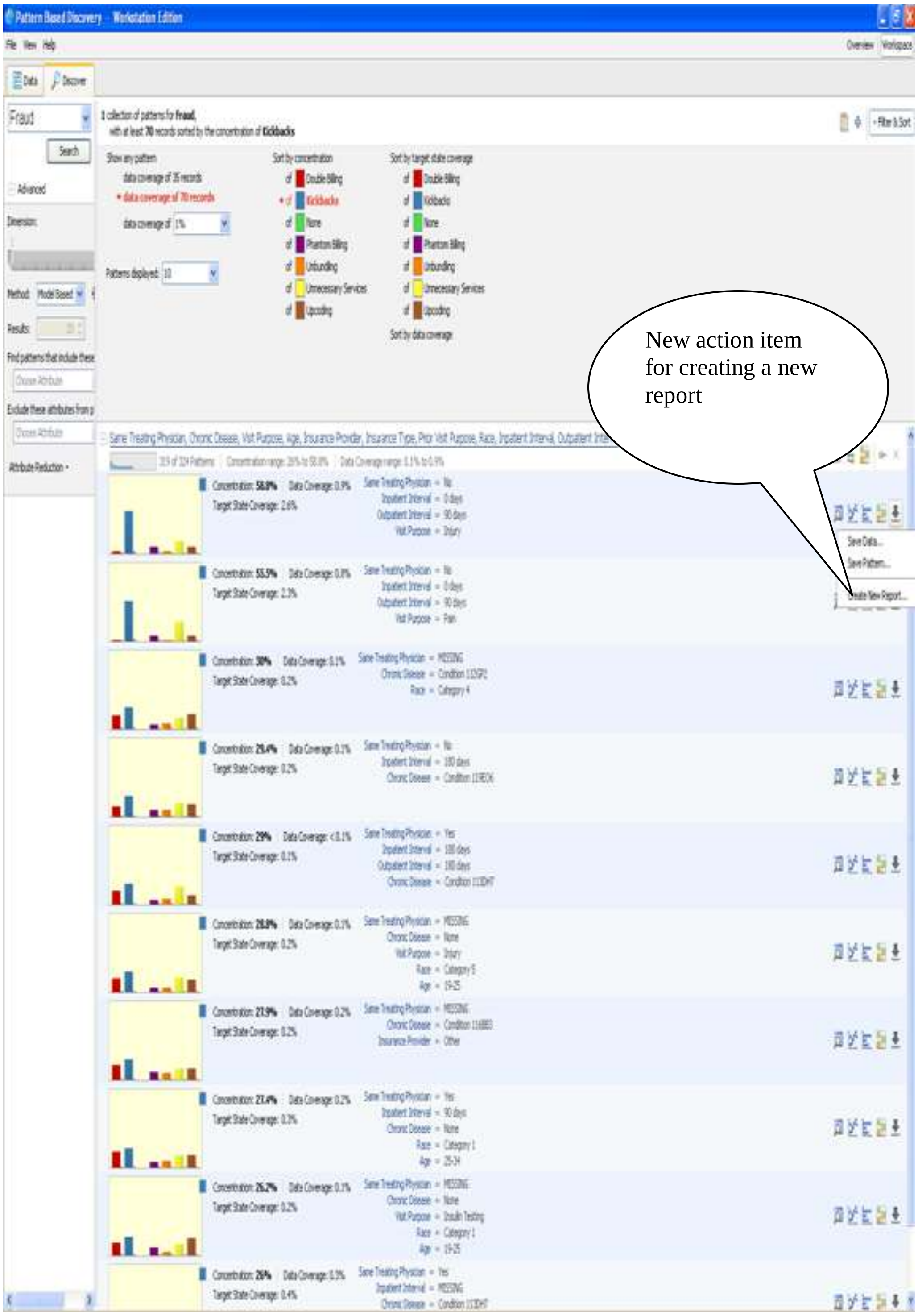
In order to allow the user to record observations of interest during the data exploration process, the Jasper Report Generator provided by JasperSoft was integrated into the Discovery product. This capability allows the user to select patterns of interest and add them to a customizable report that can be published as a pdf file at the end of the Discovery session. The workflow to realize this capability is reminiscent of adding items to a shopping cart in a consumer website that ultimately results in making a purchase.

We use the same Healthcare Fraud example from our previous quarterly report summarized below:

Healthcare Fraud Example:

It has been estimated that healthcare fraud and abuse can constitute between 3 -15% of annual healthcare expenditures in the United States. From a cost standpoint, this translates to $100-$170 billion in annual costs! Analysis of healthcare data to discover patterns that associate with different fraud types can potentially provide a proactive means for health care providers to detect fraud early on. In this example, we use a simulated data set of ~1 million patients based on an existing fraud model ("Healthcare Fraud and Abuse", Rudman et al). Six fraud types are modeled based on statistical occurrence within the nation. An additional challenge with this data set is the prevalence of MISSING data that is characteristic of healthcare data.

Figure 4.2.1 below shows the addition of a new action at the far right of the pattern summary that enables the creation of a new report.



Fraud

1 collection of patterns for **Fraud**,
with at least **70** records sorted by the concentration of **Kickbacks**

Show any pattern
data coverage of 35 records
• data coverage of 70 records
data coverage of 1%

Patterns displayed: 10

Sort by concentration
of 🟥 Double Billing
• of 🟦 Kickbacks
of 🟩 None
of 🟪 Phantom Billing
of 🟧 Unbundling
of 🟨 Unnecessary Services
of 🟫 Upcoding

Sort by target state coverage
of 🟥 Double Billing
of 🟦 Kickbacks
of 🟩 None
of 🟪 Phantom Billing
of 🟧 Unbundling
of 🟨 Unnecessary Services
of 🟫 Upcoding
Sort by data coverage

Same Treating Physician, Chronic Disease, Visit Purpose, Age, Insurance Provider, Insurance Type, Prior Visit Purpose, Race, Inpatient Interval, Outpatient Inte

319 of 324 Patterns | Concentration range: 26% to 58.8% | Data Coverage range: 0.1% to 0.9%

Concentration: **58.8%** Data Coverage: 0.9%
Target State Coverage: 2.6%
Same Treating Physician = No
Inpatient Interval = 0 days
Outpatient Interval = 90 days
Visit Purpose = Injury

Concentration: **55.5%** Data Coverage: 0.8%
Target State Coverage: 2.3%
Same Treating Physician = No
Inpatient Interval = 0 days
Outpatient Interval = 90 days
Visit Purpose = Pain

Concentration: **30%** Data Coverage: 0.1%
Target State Coverage: 0.2%
Same Treating Physician = MISSING
Chronic Disease = Condition 115GP2
Race = Category 4

Concentration: **29.4%** Data Coverage: 0.1%
Target State Coverage: 0.2%
Same Treating Physician = No
Inpatient Interval = 180 days
Chronic Disease = Condition 119EOK

Concentration: **29%** Data Coverage: <0.1%
Target State Coverage: 0.1%
Same Treating Physician = Yes
Inpatient Interval = 180 days
Outpatient Interval = 180 days
Chronic Disease = Condition 113DHT

Concentration: **28.8%** Data Coverage: 0.1%
Target State Coverage: 0.2%
Same Treating Physician = MISSING
Chronic Disease = None
Visit Purpose = Injury
Race = Category 5
Age = 19-25

Concentration: **27.9%** Data Coverage: 0.2%
Target State Coverage: 0.2%
Same Treating Physician = MISSING
Chronic Disease = Condition 116BE3
Insurance Provider = Other

Concentration: **27.4%** Data Coverage: 0.2%
Target State Coverage: 0.3%
Same Treating Physician = Yes
Inpatient Interval = 90 days
Chronic Disease = None
Race = Category 1
Age = 25-34

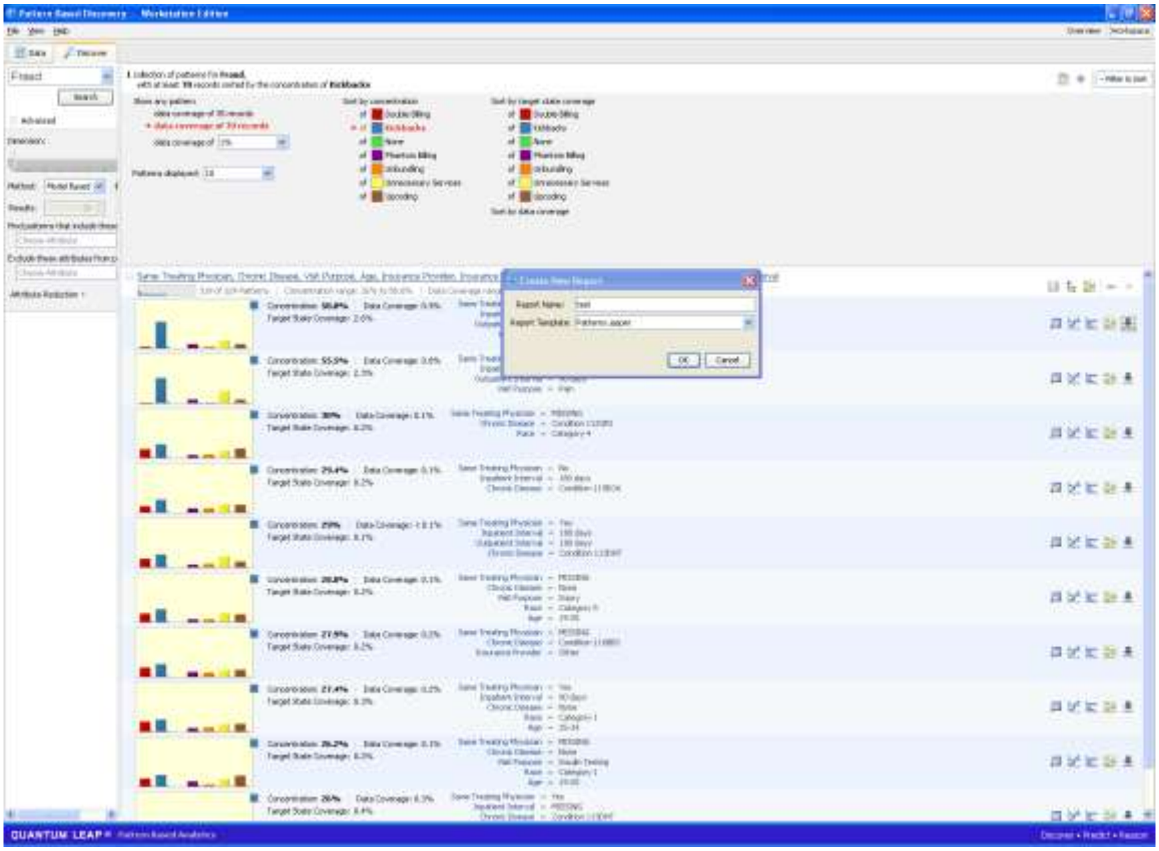
Concentration: **26.2%** Data Coverage: 0.1%
Target State Coverage: 0.2%
Same Treating Physician = MISSING
Chronic Disease = None
Visit Purpose = Insulin Testing
Race = Category 1
Age = 19-25

Concentration: **26%** Data Coverage: 0.3%
Target State Coverage: 0.4%
Same Treating Physician = Yes
Inpatient Interval = MISSING
Chronic Disease = Condition 113DHT

New action item for creating a new report

Figure 4.2.1. Addition of new action item for report creation

When the user selects “Create New Report”, an edit box to create the report appears as shown in Figure 4.2.2.

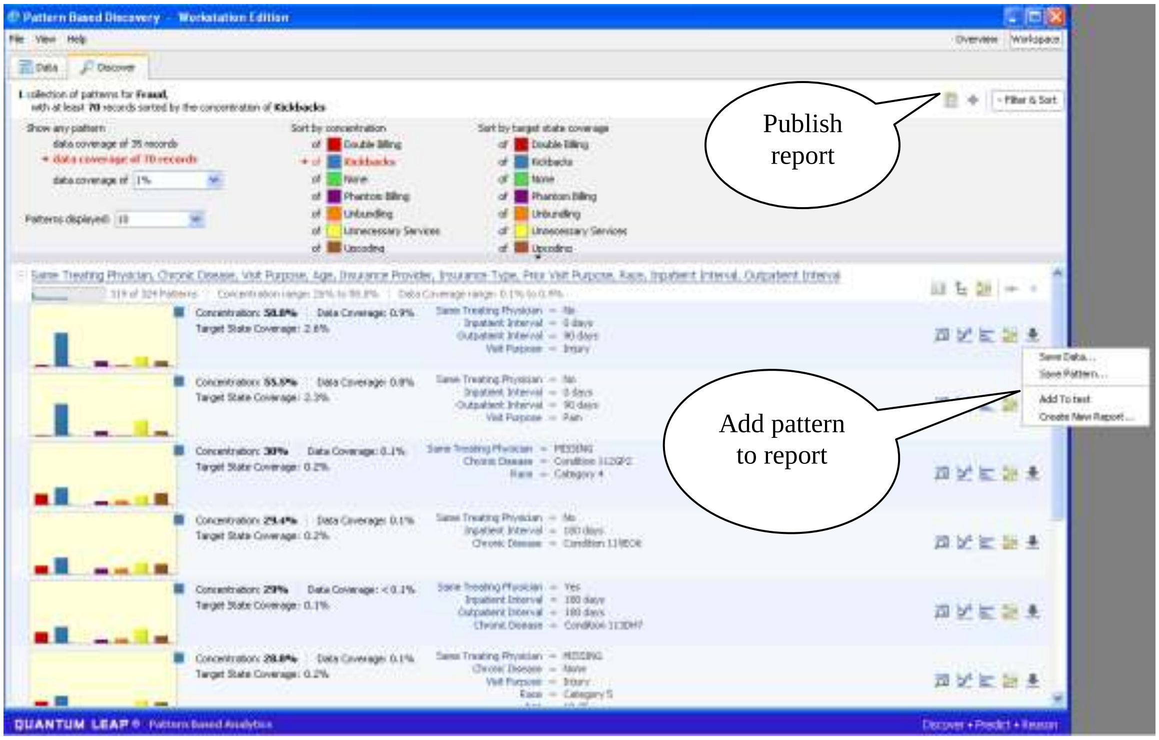


Report Name: test

Report Template: Patterns.jasper

Figure 4.2.2. Creation of a new report

After the new report has been created, a pattern can be added to the report as shown in Figure 3. The highlighted icon at the top right of Figure 3, when clicked, will display the pattern report shown in Figure 4.2.3.



1 collection of patterns for **Fraud**,
with at least **70** records sorted by the concentration of **Kickbacks**

Show any pattern
data coverage of 35 records
• data coverage of 70 records
data coverage of 1%

Patterns displayed: 10

Sort by concentration	Sort by target state coverage
of 🟥 Double Billing	of 🟥 Double Billing
• of 🟦 Kickbacks	of 🟦 Kickbacks
of 🟩 None	of 🟩 None
of 🟪 Phantom Billing	of 🟪 Phantom Billing
of 🟧 Unbundling	of 🟧 Unbundling
of 🟨 Unnecessary Services	of 🟨 Unnecessary Services
of 🟫 Upcoding	of 🟫 Upcoding

Same Treating Physician, Chronic Disease, Visit Purpose, Age, Insurance Provider, Insurance Type, Prior Visit Purpose, Race, Inpatient Interval, Outpatient Interval

319 of 324 Patterns | Concentration range: 29% to 58.8% | Data Coverage range: 0.1% to 0.9%

Concentration: **58.8%** Data Coverage: 0.9%
Target State Coverage: 2.6%
Same Treating Physician = No
Inpatient Interval = 0 days
Outpatient Interval = 90 days
Visit Purpose = Injury

Concentration: **55.5%** Data Coverage: 0.8%
Target State Coverage: 2.3%
Same Treating Physician = No
Inpatient Interval = 0 days
Outpatient Interval = 90 days
Visit Purpose = Pain

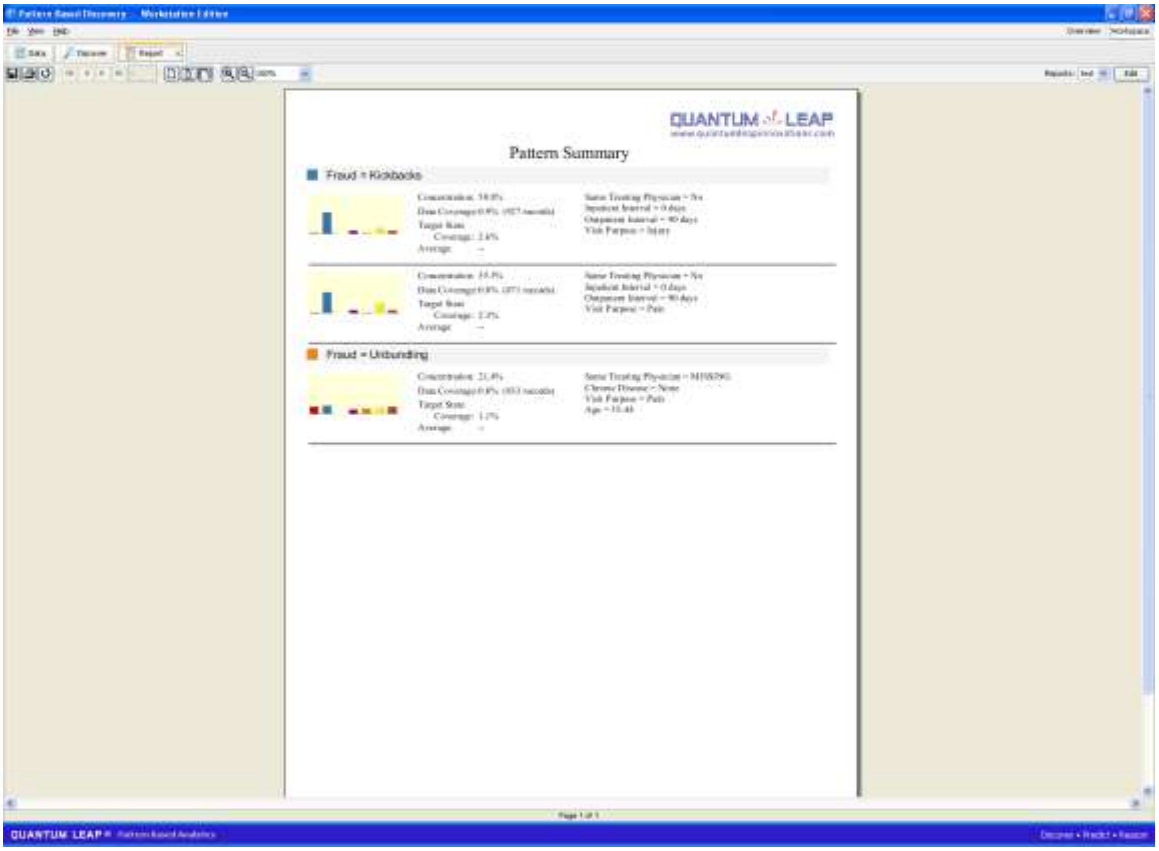
Concentration: **30%** Data Coverage: 0.1%
Target State Coverage: 0.2%
Same Treating Physician = MISSING
Chronic Disease = Condition 112GP2
Race = Category 4

Concentration: **29.4%** Data Coverage: 0.1%
Target State Coverage: 0.2%
Same Treating Physician = No
Inpatient Interval = 180 days
Chronic Disease = Condition 119EO6

Concentration: **29%** Data Coverage: < 0.1%
Target State Coverage: 0.1%
Same Treating Physician = Yes
Inpatient Interval = 180 days
Outpatient Interval = 180 days
Chronic Disease = Condition 113DH7

Concentration: **28.8%** Data Coverage: 0.1%
Target State Coverage: 0.2%
Same Treating Physician = MISSING
Chronic Disease = None
Visit Purpose = Injury
Race = Category 5

Figure 4.2.3. Adding a pattern to a report


Pattern Based Discovery - Workstation Edition
File View Help
Overview Workspace
Data Discover Report
100%
Reports: test Edit
QUANTUM LEAP
www.quantumleapinnovations.com
Page 1 of 1
QUANTUM LEAP® Pattern Based Analytics
Discover • Predict • Reason

Pattern Summary

🟦 Fraud = Kickbacks

Concentration: 58.8%
Data Coverage: 0.9% (127 records)
Target State
Coverage: 2.6%
Average: --
Same Treating Physician = No
Inpatient Interval = 0 days
Outpatient Interval = 90 days
Visit Purpose = Injury

Concentration: 55.5%
Data Coverage: 0.8% (171 records)
Target State
Coverage: 2.3%
Average: --
Same Treating Physician = No
Inpatient Interval = 0 days
Outpatient Interval = 90 days
Visit Purpose = Pain

🟧 Fraud = Unbundling

Concentration: 21.4%
Data Coverage: 0.8% (103 records)
Target State
Coverage: 1.2%
Average: --
Same Treating Physician = MISSING
Chronic Disease = None
Visit Purpose = Pain
Age = 31-40

Figure 4.2.4. Example of a generated report

4.2.2 Addition of Spreadsheet capability:

In order to facilitate analysis of a data subset associated with a pattern, the data contained within a pattern can be automatically copied to a spreadsheet. This capability was implemented using JComponentPack v3.3.4. In Figure 4.2.5, the scratchpad icon on the top right of the pattern data table copies the data table into a spread sheet as shown in Figure 4.2.6.

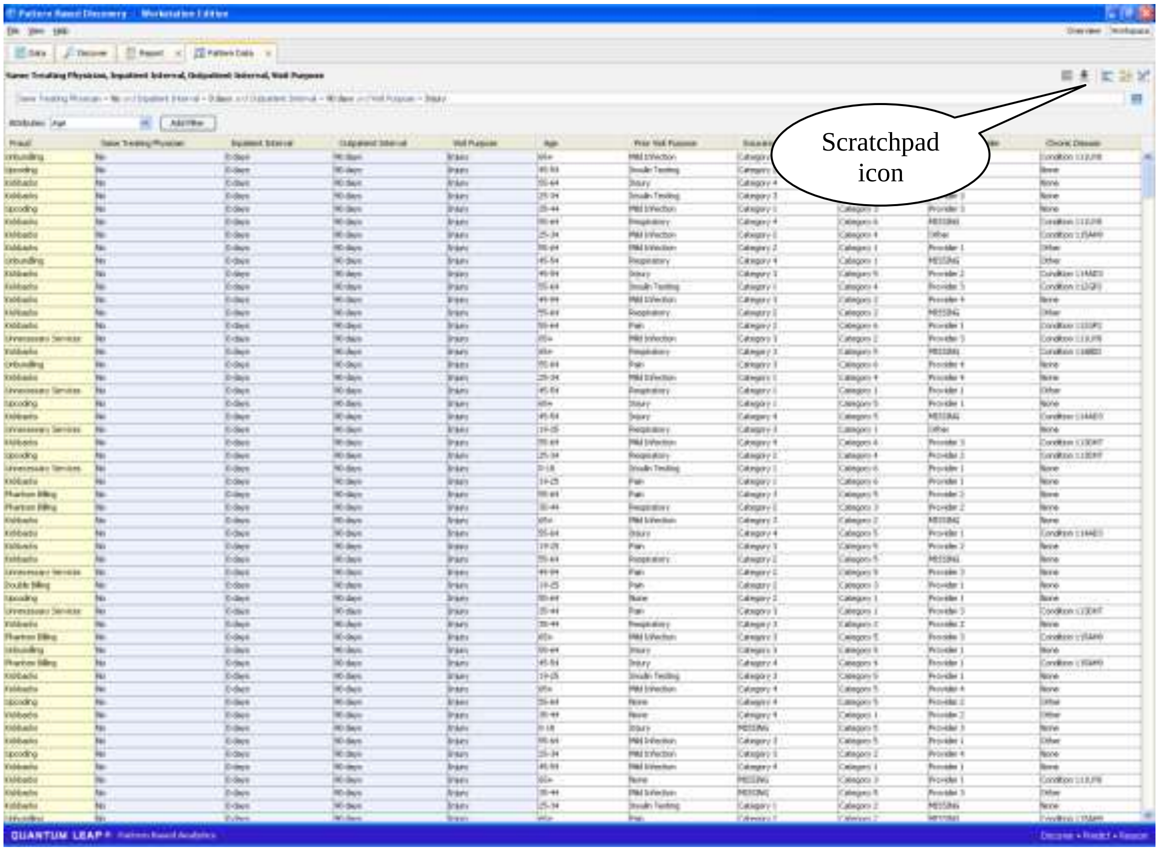


Figure 4.2.5. Scratchpad icon on top right of Pattern Data Table

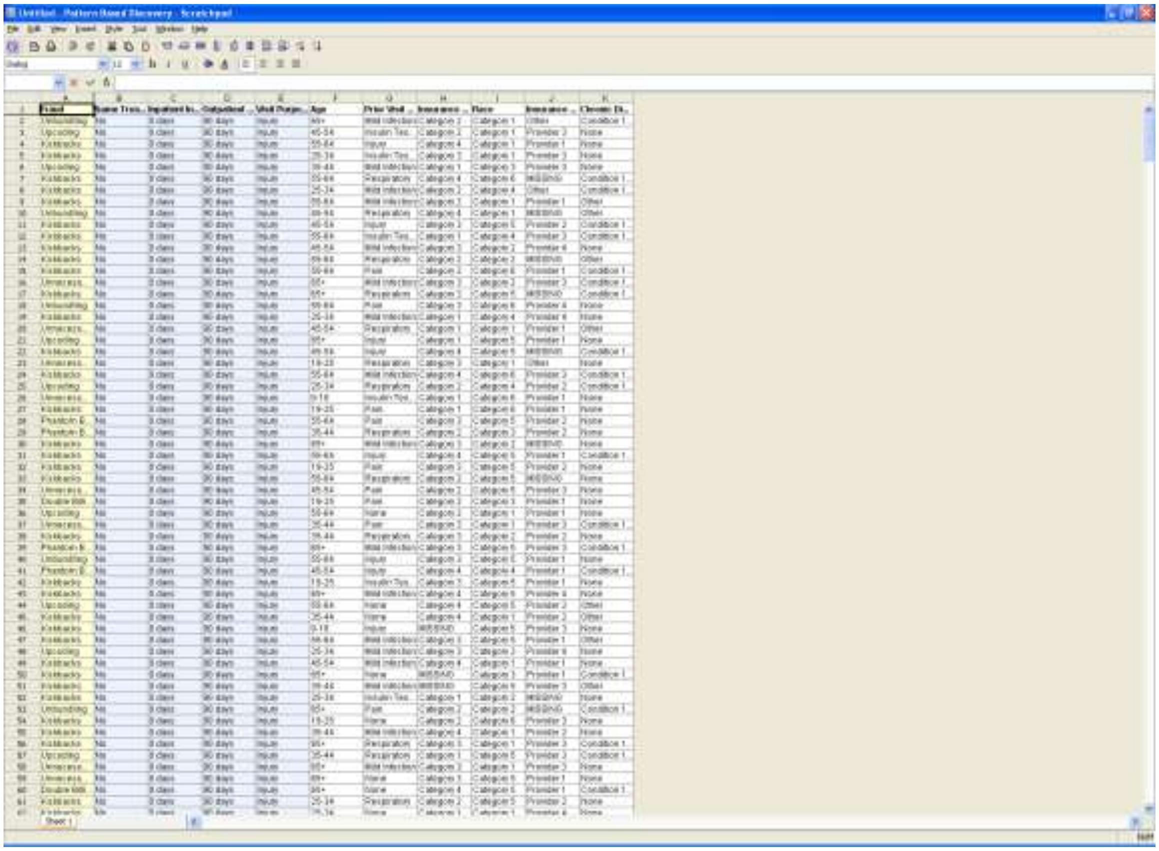


Figure 4.2.6. Spreadsheet copy of Pattern Data Table

The spreadsheet implemented within Discovery allows the user to enter formulae, insert columns and perform several of the most commonly used functions within Excel.

4.2.3 Implementation of stochastic decision trees for Pattern Based Discovery

The search engine was refined to include ensembles of stochastic decision trees to increase the diversity of the search through the data environment. Patterns collected from the ensemble of trees can increase the number of useful patterns isolated within multivariate data sets. This can also provide a natural bridge to Pattern Based Prediction described below.

4.3 PATTERN BASED PREDICTION

Quantum Leap has also implemented ensembles of stochastic decision trees as a basis for Pattern Based Prediction. The primary motivation is to use the same modeling paradigm used in Discovery for implementing Pattern Based Prediction. This allows a seamless integration of a Business Intelligence capability defined by Discovery with a Business Analytics capability defined by Prediction so that users can easily go back and forth between BI and BA.

Pattern Based Prediction Example KDD Cup 2008 - Breast Cancer Identification

Motivation:

A breast cancer screen typically consists of 4 X-ray images; 2 images of each breast from different directions (these views are called MLO and CC). Thus, most (but not all) patients would have MLO and CC images of both their breasts, giving a total of 4 images per patient. For the purposes of the KDD Cup, each image is represented by several candidates (see stage 1 above). For each candidate, we provide the image ID and the patient ID, (x,y) location, several features, and a class label indicating whether or not it is malignant. We provide features computed from several standard image processing algorithms – 117 in all – but due to confidentiality reasons we are unable to provide some additional proprietary features. The labels indicate whether a candidate is malignant or benign (based on either a radiologist's interpretation or a biopsy or both). Note that several candidates can correspond to the same lesion.

The training set consists of 50,563 data records with 118 features including all 4 views for each patient plus the target feature. The proportion of malignant tumors to benign Tumors is 1:151, making this a "needle in a haystack" type problem.

Figure 4.3.1 shows the Data Screen for Pattern Based Prediction that has a similar look and feel to the corresponding Data Screen for Pattern Based Discovery. On the bottom left of Figure 4.3.1, we note the presence of three tabs: Data, Discovery and Prediction that allows the user to navigate easily between the three functions. The histogram on the right shows the distribution of malignant tumors in this data set, reflecting the needle in a haystack behavior described above.

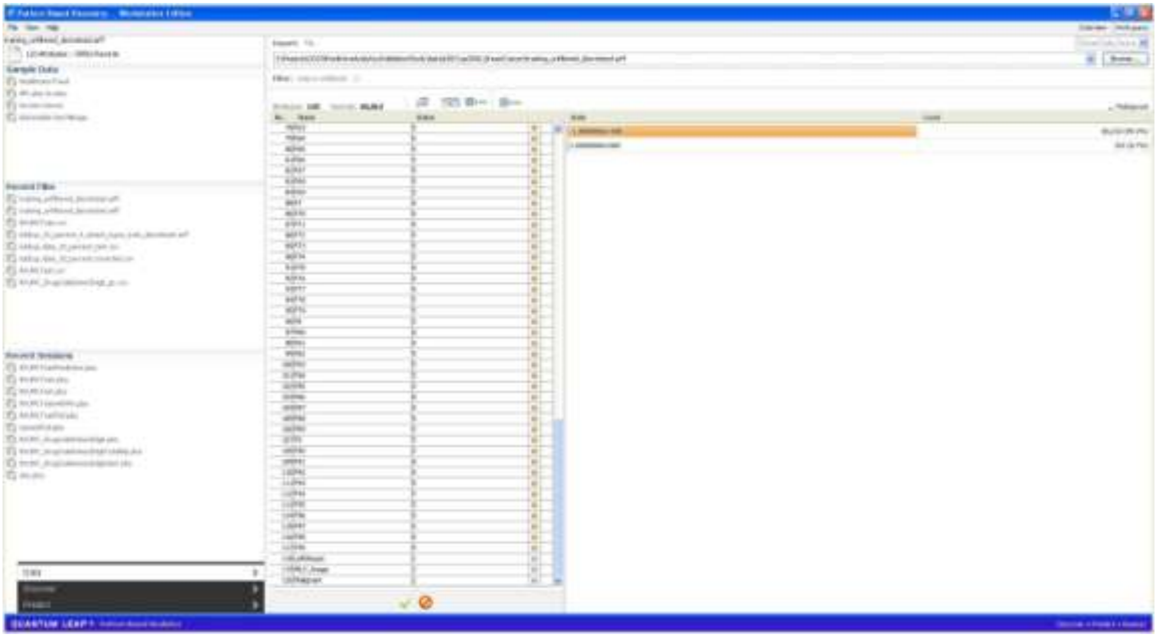


Figure 4.3.1 Data screen for Prediction

Figure 4.3.2 shows the Prediction Screen where the modeling conditions are summarized on the left. 20 stochastic models were built on a training set that consists of 75% of the data with the remaining 25% of the data being used as the test set to evaluate the predictive model. The plot on the right represents a standard ROC curve that plots the True Positive Rate versus the False Positive Rate for malignant tumor detection. The area under the ROC curve would be 0.5 for a model that guesses randomly and 1.0 for a perfect model. In this example, the area under the ROC curve is 0.95, indicating the high quality of the model even for this very difficult problem. Summary statistics for both the training data used to build the model as well as the test data used to evaluate the model are displayed below the ROC curve.

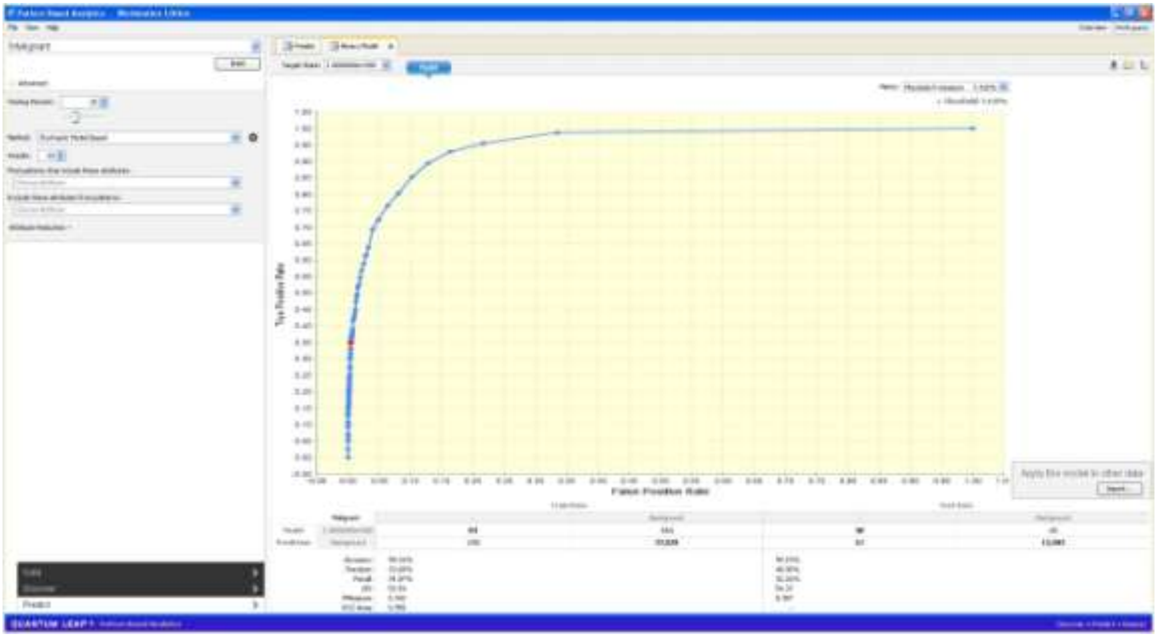


Figure 4.3.2 Prediction Screen